공학석사 학위논문

토픽모델링과 사회연결망분석을 활용한 '사물인터넷' 연구동향 분석

Research Trends Analysis for 'Internet of Things' Based on Topic Modeling and Network Analysis

2016년 2월

서울과학기술대학교 일반대학원

데이터사이언스학과

김 준 홍

토픽모델링과 사회연결망분석을 활용한 '사물인터넷' 연구동향 분석

Research Trends Analysis for 'Internet of Things' Based on Topic Modeling and Network Analysis

지도교수 이원영

이 논문을 공학석사 학위논문으로 제출함

2016년 2월

서울과학기술대학교 일반대학원

데이터사이언스학과

김 준 홍

김준홍의 공학석사 학위논문을 인준함

2016년 2월

심사위원장 조남욱 (인)

심사위원 이원영 (인)

심사위원 이윤호 (인)

목 차

요 약

제 목 : 토픽모델링과 사회연결망 분석을 활용한 ‘사물인터넷’ 연구동향 분석

본 연구에서는 미래 주요 기술 중의 하나로 각광받고 있는 ‘사물인터넷’ 분야의 연구 동향을 토픽 모델링과 사회연결망 분석 기법을 이용하여 분석하였다. 이를 위해 2003년부터 2015년까지 ‘사물인터넷’과 관련된 국제학술지 및 학술대회 논문 6,634편을 대상으로 출판연도-국가별 빈도 분석 및 네트워크 중심성 분석을 통해 국가 간 협업 현황을 분석한 결과, 출간 수 기준 상위 5개의 나라는 China, United States, Germany, Italy, United Kingdom으로 산출 되었으며, 국가별 공동연구에서 가장 중심이 되는 나라는 매개 중심성 기준으로 보았을 때, United Kingdom, United States, China, Germany, Spain순으로 밝혀졌다. 이는 국가별 공동 연구의 중개자의 순위가 같다고 할 수 있다.

제목과 저자 키워드를 기반으로 하는 단어 네트워크 그래프 구축을 통해 8가지의 사물인터넷 핵심 연구 분야를 도출하였다. 또한, 각 논문의 ‘제목, 초록, 연구문서 키워드’에 대한 Latent Dirichlet allocation을 사용한 토픽모델 기법 적용을 통해 ‘사물인터넷’ 연구 분야 20가지 토픽을 도출한 후 시간별 전체 연구비중으로 산출되는 회귀식의 증가 감소 추세와 회귀계수의 p-value를 통해 Hot/Warm/Cool/Cold Topic으로 분할하였다. 연구 비중이 증가 추세에 있으며 통계적으로 유의미한 Hot Topic으로는 Security & Privacy, Smart Home/City/Grid, Big Data Processing, Object/Signal Detection, Routing Algorithm & Protocol으로 분석되었고, 연구비중이 감소 추세에 있으며 통계적으로 유의미한 Cold Topic으로는 RFID, Virtual reality & User Experience, Supply Chain Management로 분석 되었다. 또한 토픽모델로 산출된 20가지 토픽과 제목 및 저자 키워드 기반 네트워크에서 도출된 8가지 연구 분야와의 관계를 분석하였다.

본 연구 결과는 사물인터넷 분야의 전반적인 연구 동향 및 향후 전략을 수립하는데 기초 자료로 활용될 수 있을 것으로 기대한다.

표 목 차

그림목차

수식목차

Ⅰ. 서 론

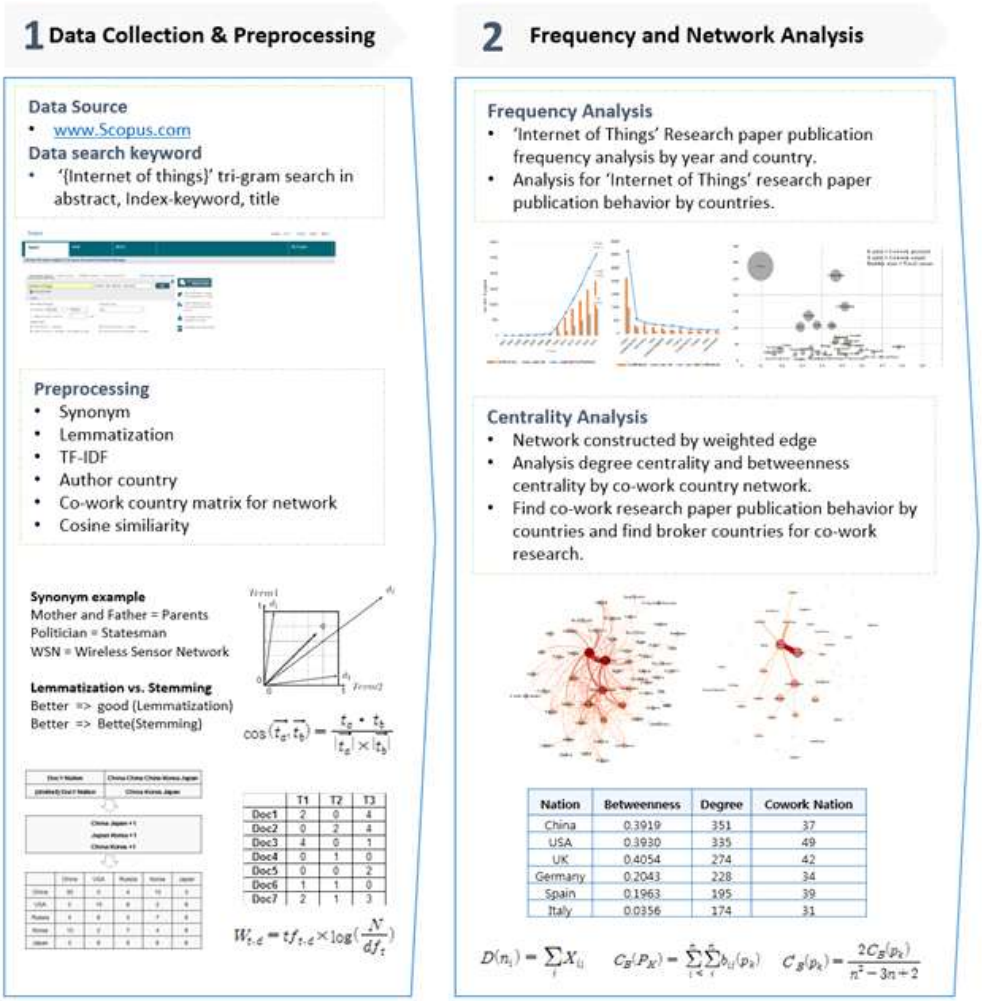
사물 인터넷(Internet of Things)은 1999년의 영국의 Kevin Ashton에 의해 처음 제안된 용어로써, 사물 간 통신을 위한 향상된 장치 및 시스템 그리고 이러한 네트워크상에서 발생하는 다양한 서비스를 모두 망라하는 개념으로 사용되고 있다[1]. 1990년대 월드와이드웹(world wide web; WWW)의 등장과 2000년대 후반 스마트폰의 대중화로 인해 인간과 사물의 네트워크에 대한 연결성이 급격히 증가하고 있다. 이러한 초연결 사회에서 사물인터넷은 고유 주소를 보유한 객체들의 상호작용을 바탕으로 일상생활뿐만 아니라 자동차 및 산업제조, 물류, 비즈니스/프로세스 관리, 지능형 교통 등의 분야에서 혁신을 주도할 패러다임으로 급부상 하고 있다. 컨설팅 그룹 맥킨지에서는 사물인터넷을 향후 2025년 까지 인류 산업에 큰 영향을 주는 기술 중 하나로 선정하였으며[2], 시장 조사기관인 가트너에서는 현재 사물인터넷이 기술 순환 주기 중 기대치가 정점에 이르렀다고 평가하고 있다[3].

본 연구에서 다룰 '사물 인터넷 관련 연구 트렌드 분석'에 대한 기존의 선행 연구들을 살펴보면 다음과 같다. Atzori et al. (2010)은 사물인터넷 관련 연구논문들에 대해 전문가 지식(domain knowledge)을 이용하여 사물인터넷에서 사용되는 기술과 적용되고 있는 상위분야 5가지 및 세부분야 19가지에 대하여 서술하였다[4]. Xu et al. (2014)은 사물인터넷에 대한 88편의 연구논문에 대해 전문가 지식을 바탕으로 사물인터넷을 위한 서비스 기반 아키텍처를 4가지 층으로 분류한 뒤, 각각에 대한 연구와 핵심기술을 요약하고 그 트렌드를 해석 하였다[5]. 또한 Whitmore et al. (2015)은 사물인터넷 관련 127편의 연구논문을 대상으로 전문가 지식을 이용하여 사물인터넷 관련 연구를 6가지의 대분류로 분류하였으며 이 중 3가지 대분류를 11가지 중분류로, 다시 3가지 중분류를 9가지 소분류로 분류 하여 각 분류별 연구결과를 요약 하였다[6].

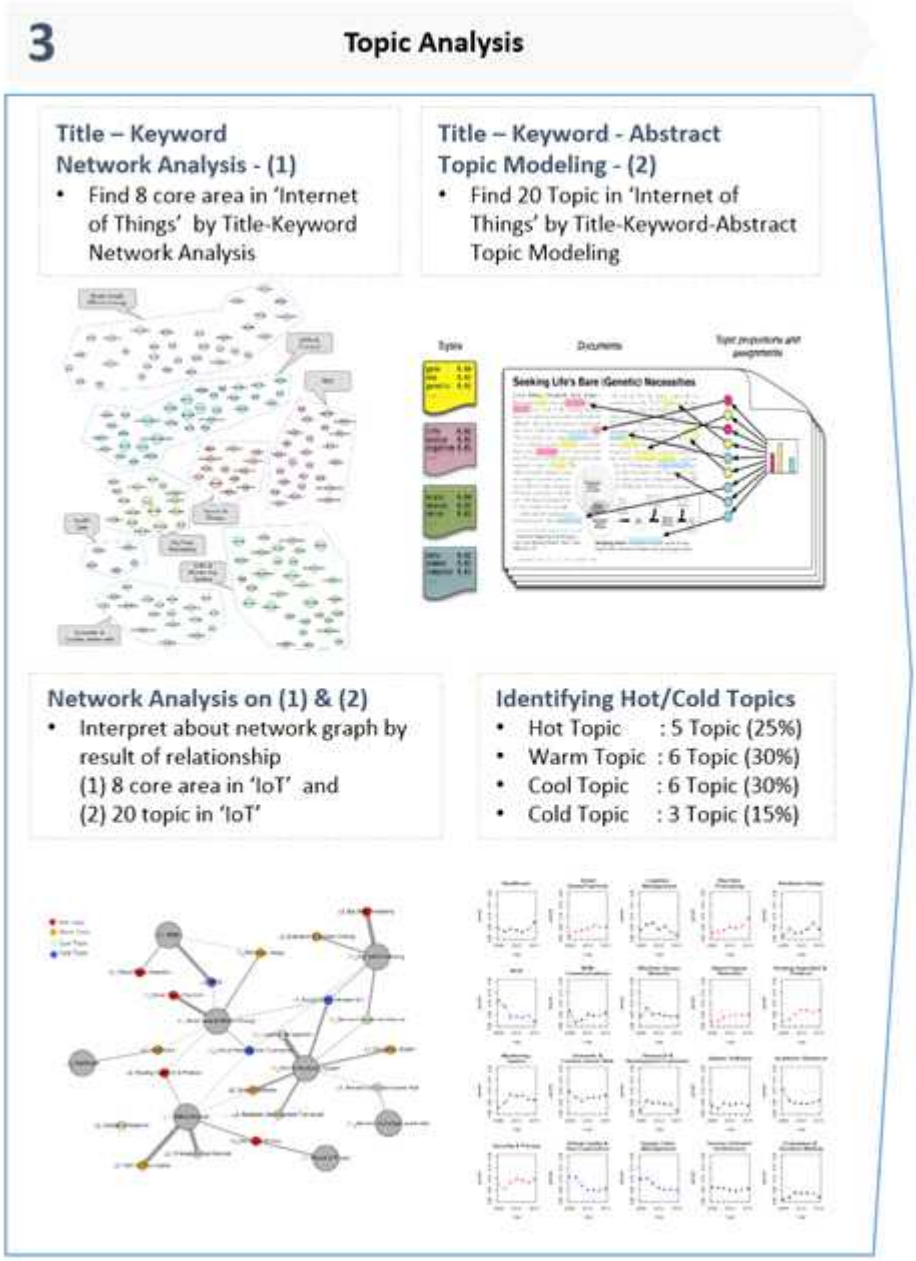
이와 같이 사물인터넷 관련 문헌 분석은 대부분 전문가의 지식에 크게 의존하고 있으며 대용량의 연구논문을 대상으로 정량적인 분석을 수행한 연구는 거의 없는 실정이다. 따라서 본 연구는 현재 출간된 사물인터넷 관련 연구논문의 제목, 초록, 소속, 연도, 저자 제공 키워드 등의 정보를 기반으로 bag of words (BOW)방식의 문서 집합(corpus)을 구축하고 이를 바탕으로 빈도분석, 네트워크 중심성 분석 및 토픽모델링을 통해 사물인터넷 관련 공동연구 국가 특징 분석, 핵심 키워드 도출, 대표 연구주제 판별, 주제별 트렌드 분석 등을 수행하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 데이터 수집과 연구에서 사용할 방법론을 간략히 소개한다. 3장에서는 연도-국가별 빈도분석, 국가별 공동연구 네트워크, 키워드 네트워크를 통한 사물인터넷 핵심 연구분야 도출, 토픽 모델링을 통한 20가지의 연구분야를 도출하고 제목 및 저자 키워드 기반 네트워크에서 도출된 8 가지 연구분야와의 관계를 분석한다. 4장에서는 본 연구의 결론을 서술한다.

본 연구 수행 단계는 <Fig. 1.1>과 <Fig .1.2>와 같다. 먼저 '사물 인터넷' 관련 연구논문을 수집한 뒤 전처리 과정으로 저자 소속에 따른 국가 변환, 동의어 및 다의어 처리, Lemmatization, TF-IDF, 코사인 유사도 변환 등을 수행한 후 연도-국가별 빈도분석, 공동연구 국가의 네트워크 중심성 지표 분석, 국가별 연구문서 출간성향 분석, 키워드 네트워크를 통한 연구분야 도출, 토픽 모델링을 통한 주제 탐색 및 주제별 트렌드 분석, 토픽모델링 주제와 키워드 네트워크 결합 분석을 수행한다.



<Fig. 1.1> Research framework



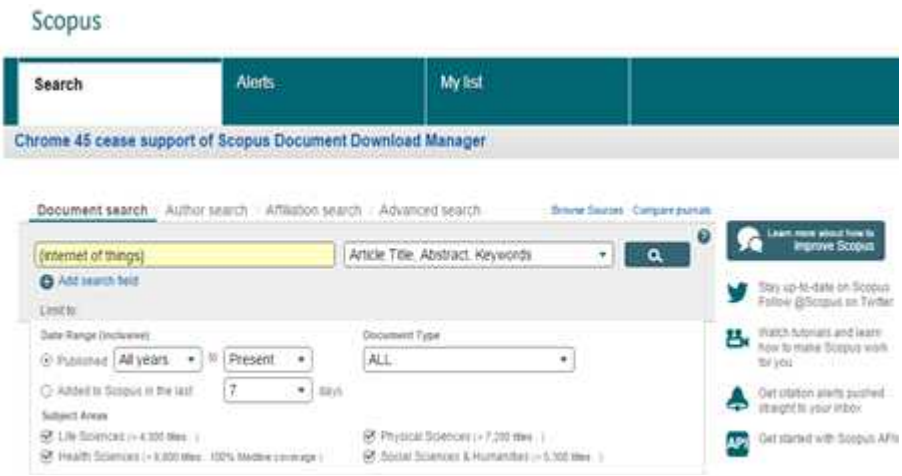
<Fig. 1.1> Research framework (Cont'd)

Ⅱ. 연구방법론

Ⅱ장에서는 데이터 수집에 대한 상세한 설명과 수집결과, 해당 연구에서 사용한 방법론에 대하여 설명하고 그 이유에 대하여 서술한다.

2.1. 데이터 수집 및 전처리

본 연구에서 사용된 연구논문은 Scopus (www.scopus.com) 사이트를 통해 수집하였다. 제목, 초록, 저자 키워드 중 하나 이상의 항목에서 'internet of things' 라는 3음절(tri-gram) 단어가 사용된 모든 학술지(journal) 및 학술대회(conference) 논문을 분석 대상으로 하였으며 총 6,634개의 논문에 대한 41가지 메타 정보(저자, 연도, 소속기관, 국가 등)를 수집하였다. Scopus의 검색 화면은 <Fig. 2.1>과 같고 수집된 데이터 요약은 [Table 2.1]과 같다.



<Fig. 2.1> Scopus(www.scopus.com)

[Table 2.1] 수집 데이터 요약

문서의 갯수	6,634 documents
변수의 갯수	41 columns
수집처	www.scopus.com
키워드	{internet of things} (tri-gram search)
문서타입	Journal paper or Conference Paper
검색대상	Title, Abstract, Keywords

각 논문의 제목, 초록 및 저자 키워드는 <Eq. 2.1>에 나타난 바와 같이 특정 문서에 대한 개별 단어의 중요도를 고려한 단어 빈도-역문서 빈도(Term Frequency - Inverse Document Frequency; 이하 TF-IDF) 방식을 활용하여 벡터 형태로 표현하였다.

$$W_{t,d} = tf_{t,d} \times \log_e(\frac{N}{df_t})$$

<Eq. 2.1> TF-IDF equation

TF-IDF의 수치가 크다는 뜻은 특정 단어(Term)가 문서를 대표할 확률이 높다는 것을 의미한다. ‘t’는 단어(Term)의 번호 ‘d’는 문서(Document)의 번호 ‘N’은 문서(Document)의 총 갯수를 의미한다. ‘tf’(Term Frequency)는 특정 문서(Document)에서 특정 단어(Term)이 몇 번의 빈도로 발생했는지를 의미한다. ‘df’(Document Frequency)는 전체 문서(Document)에서 특정 단어(Term)가 관찰된 문서(Document)의 횟수이다.

예를 들어 3개의 문서에서 Keyword A,B,C가 존재하며 각 문서에서 키워드의 존재여부를 산출 후 이진횟수 행렬로 변환 하였다고 가정 하였을 때, [Table 2.2]처럼 이를 Document Term matrix로 변환 할 수 있다.

[Table 2.2] Document Term matrix (TF)

	Keyword A	Keyword B	Keyword C
Document1	1	1	0
Document2	1	1	0
Document3	1	0	1

이것을 다시 TF(Term Frequency)를 TF-IDF로 변환하게 되면 [Table 2.3]과 같고 ‘Keyword C’의 경우는 3개의 문서중 하나만 나타났기 때문에 IDF(Inverse Document Frequency)에 대한 가중치가 크게 작용하게 되며 이는 해당 Keyword C가 문서 분류로써 중요한 키워드라는 것을 의미한다. 또한, ‘Keyword A’는 모든 문서에서 발견되는 키워드로써 중요성이 없으며 수식상 0으로 출력되게 된다.

[Table 2.3] Document Term matrix (TF-IDF)

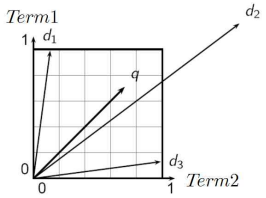
'	Keyword A	Keyword B	Keyword C
Document1	0	0.405	0
Document2	0	0.405	0
Document3	0	0	1.098

문서 간 혹은 단어 간 유사도 측정이 필요한 경우, 본 연구에서는 <Eq. 2.2>과 같이 정의된 코사인 유사도를 사용하였다. A의 문서와 B의 문서의 키워드 백터를 기반으로 코사인 유사도 값을 구할 수 있다. 코사인 유사도는 텍스트 마이닝 분야에서 가장 널리 사용되는 유사도 측정 기법의 하나로써 유클리디언 거리와는 달리 단어의 절대적인 사용 빈도가 아닌 전체 문서에서 단어의 사용 비율을 고려함으로써 서로 다른 길이의 문서들을 보다 효과적으로 비교할 수 있다는 장점이 있다.

$$\cos(\theta) = \frac{\overrightarrow{A} \bullet \overrightarrow{B}}{\|\overrightarrow{A}\| \|\overrightarrow{B}\|}$$

<Eq. 2.2> Cosine similarity equation

<Fig. 2.2>와 같이 'q' 벡터(Query)와 'd2' 벡터(Document2)의 상황처럼 Euclidean distance 사용 시 Term-Frequency의 비율은 비슷하나 횟수의 대소관계가 커지게 되면 Similarity가 작아지게 되는 상황을 각도 지표인 Cosine값으로 대체할 수 있다.



<Fig. 2.2> Document 1,2,3와 Query(New Document) 의
2차원 Term vector space

수집된 논문은 모두 영어로 작성되어 있는데 영어 단어는 동일한 단어가 문법적 역할에 따라 여러 형태로 변형되어 사용된다. 이러한 문제를 해결하고 단어의 기본형을 추출하기 위한 방식으로 Stemming과 Lemmatization이 존재한다. Stemming은 단어의 근원(root)을 찾아주는 방식이고 Lemmatization은 단어의 기본형(lemma)을 찾아주는 방식이다. 본 연구에서는 Lemmatization을 통해 단어 처리를 수행하였는데 그 이유는 Lemmatization은 Stemming의 문제점 중 하나인 두 개 이상의 다른 단어가 동일한 Stem으로 추출되는 것을 방지할 수 있으며 이에 더하여 품사 정보 또한 유지할 수 있기 때문이다.

2.2. 사회 연결망 기반의 중심성 분석

사회 연결망 분석은 네트워크 및 그래프 이론의 사용을 통해 사회 및 현상의 구조를 모사하고 그 특징을 분석하는 데 사용되는 기법이다[7]. 이 중에서 중심성(centrality)은 네트워크 안에서 중요한 역할을 하는 개체(노드)가 무엇인지를 판별해주는 지표로써 구축된 네트워크를 해석하는 데 있어 중요한 역할을 하게 된다[8]. 본 연구에서는 국가 간 협업 네트워크 현황을 분석하기 위해 연결 중심성(degree centrality)과 매개 중심성(betweenness centrality)을 측정하고 그 결과를 해석하였다. 연결 중심성은 <Eq. 2.3>와 같이 계산되는데 i는 노드의 번호이며 j는 전체 노드에서 i를 제외한 나머지 노드들이다. 또한 X_{ij}는 i번째 노드와 j번째 노드 사이에 연결된 호(edge)의 크기이다. 연결 중심성은 호의 가중치를 사용하지 않은 네트워크에서 하나의 노드가 몇 개의 인접한 노드와 연결되어 있는지를 나타내며 가중된 호를 이용한 네트워크에서는 하나의 노드에 연결 되어 있는 호들의 가중치 총합을 나타낸다.

$$D(n_i) = \sum_j X_{ij}$$

<Eq. 2.3> Degree Centrality Equation

매개 중심성은 <Eq. 2.4>과 같이 계산되며 네트워크상의 두 노드 사이의 최단거리에 노드 A가 위치하는 비중이 높을수록 매개중심성이 높아지게 된다[9]. σ_{jk}는 노드 j에서 k까지의 최단경로 수이고 $\sigma_{jk}(i)$는 노드 j에서 k까지의 최단경로 중 노드 i를 지나는 경로의 수이다. 매개중심성이 높은 노드는 네트워크에서 그룹과 그룹을 이어주는 매개체 역할을 하게 된다.

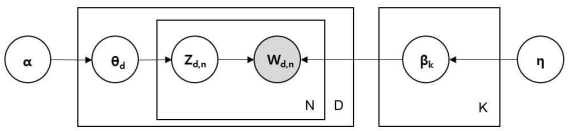
$$B(i) = \sum_{i \neq j \neq k} \frac{\sigma_{jk}(i)}{\sigma_{jk}}$$

<Eq. 2.4> Betweenness Centrality Equation

본 연구와 같이 특정 주제 키워드를 이용한 네트워크 분석을 수행한 연구들로는 'big data'를 검색어로 연구논문들을 수집한 후 핵심 키워드와 국가 간의 관계에 대한 네트워크 분석을 한 연구[10], 특허 분야 에서 네트워크 분석을 통하여 LED와 wireless broadband 영역에서 함께 사용한 특허를 분석한 연구[11], 알츠하이머 질병에 대한 문서를 바탕으로 토픽모델 및 저자 네트워크 분석을 한 연구 등이 있다[12].

2.3. 토픽 모델

토픽 모델(topic model)은 다수의 문서로 구성된 문서 집합의 기저에 존재하는 주제를 파악하는 데 사용되는 기법이다[13]. 토픽 모델은 문서 집합으로부터 특정 개수의 주제를 판별할 수 있으며, 전체 단어들의 발생 확률 분포는 각 주제마다 다르게 추정이 된다. 또한 각 문서는 여러 개의 주제가 공존하며 각 주제들의 비율은 그 문서에 사용된 단어들이 속한 주제를 바탕으로 추정할 수 있다. 본 연구에서는 토픽 모델을 수행하기 위해 잠재 디리클레 할당(latent dirichlet allocation; 이하 LDA) 기법을 이용한다. LDA는 <Fig. 2.3>과 같이 문서가 생성되는 프로세스를 확률과정으로 표현하는 기법으로써, 관측된 단어인 Wd,n과 문서별 주제 분포 파라미터 α 및 토픽별 단어 분포 파라미터 η를 통해 주제별 단어의 확률 분포 β_k와 문서별 주제 분포 θ_d를 추정하는 기법이다[14].



<Fig. 2.3> LDA document generation process

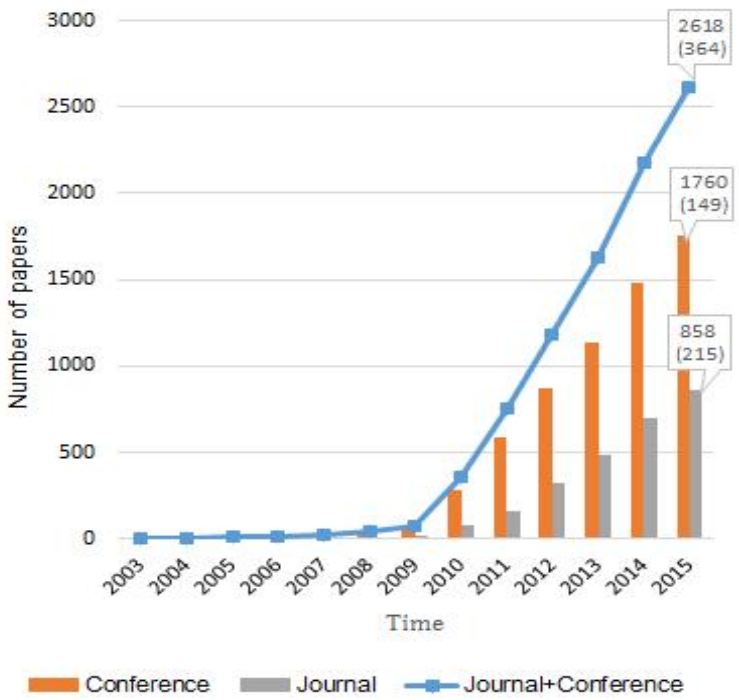
본 연구와 같이 토픽 모델을 사용하여 문서를 토픽별로 분류하고 분석한 대표적인 선행 연구로는 생물정보학 관련 문서를 토픽모델 및 저자 네트워크 등을 이용하여 분석한 연구[15], 인지분야에 대한 연구 트렌드 분석을 토픽 모델링을 통하여 확인한 연구[16], DBLP의 학술대회 데이터를 중심으로 토픽 모델링을 통해 컴퓨터 공학 및 정보학 분야의 연구동향을 분석한 연구 등이 있다[17].

Ⅲ. 분석 결과

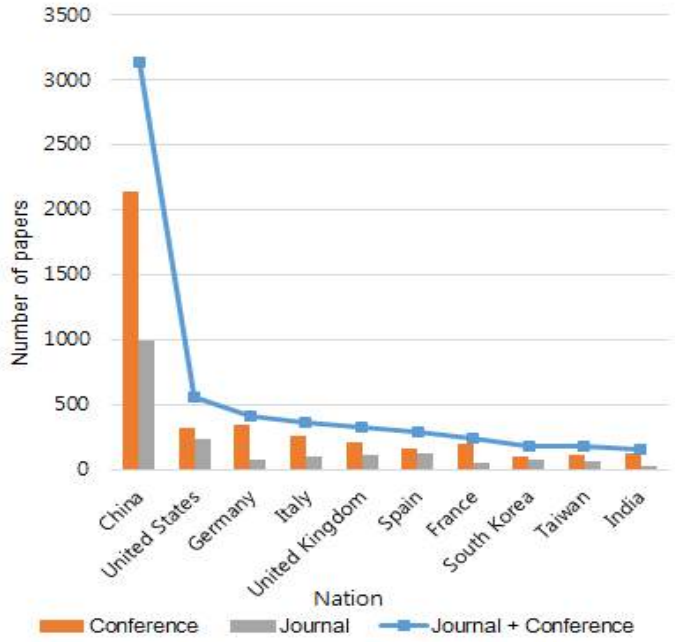
Ⅲ장에서는 수집한 데이터를 전처리후 Ⅱ장에서 언급한 분석 방법론 등을 통해 '사물인터넷' 연구동향에 대하여 분석한 결과를 서술한다.

3.1. 기초 빈도 분석

<Fig. 3.1>은 사물 인터넷 관련 학술지 및 학술대회 게재 및 발표 논문의 출판 추이를 나타낸 것이다. 이를 통해 사물 인터넷 분야의 연구논문 출판은 학술지와 학술대회에서 공통적으로 2010년 이후 급격히 증가하는 추세를 보이고 있음을 알 수 있다. 2015년의 경우는 분석시점에서 2015년이 아직 끝나지 않았기 때문에 '예측값(실제값)'으로 표시되어 있고 2011년부터 2014년까지의 수치를 이용한 회귀 모형을 통해 예상된 수치를 2015년 값으로 표현하였다.



<Fig. 3.1> Number of published papers



<Fig. 3.2> Top 10 countries in terms of the number of published papers

<Fig. 3.2>는 저자 소속 국가별 학술지 및 학술대회 연구논문 수 상위 10개를 나타낸 것이다. 사물 인터넷 관련 분야에서는 China가 압도적인 숫자의 논문(3131편)을 출판하여 출판 수 기준 1위를 차지하고 있으며 United States > Germany > Italy > United Kingdom순으로 순위가 나타나는 것을 알 수 있다. 학술지 출판 수 기준 상위 5개국은 China > United States > Spain > United Kingdom > Italy 순으로 나타났으며, 학술대회 출판 수 기준 상위 5개국은 China > Germany > United States > Italy > United Kingdom 순으로 나타났다. 우리나라의 경우 178편의 연구논문을 출판하여 총 출판 수 기준으로 8위를 차지하고 있으며, 중국을 제외한 아시아 국가 중 가장 활발하게 사물인터넷 관련 연구를 수행하는 것으로 확인되었다. 한 가지 흥미로운 사항은 Spain의 경우 학술지 기준으로는 3위를 차지하나 학술대회 기준으로는 5위 내에 존재하지 않으며 반대로 Germany의 경우 학술대회 기준 2위임에도 불구하고 학술지 기준으로는 5위 내에 존재하지 않는 것으로 나타났다. 상위 1~5위의 각 케이스는 [Table 3.1]과 같다.

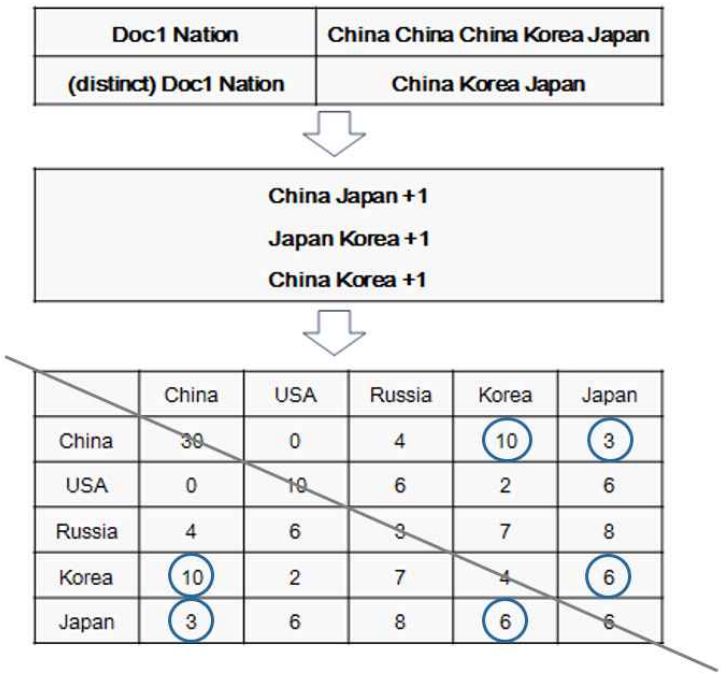
[Table 3.1] 3가지 경우에 따른 나라별 출간 횟수 순위

	1위	2위	3위	4위	5위
Journal+Conference	China	USA	Germany	Italy	UK
Conference	China	Germany	USA	Italy	UK
Journal	China	USA	Spain	UK	Italy

3.2. 국가별 협업 연구 현황 분석

국가별 공동연구 네트워크를 구축하기 위하여 <Fig. 3.3>과 같은 방식을 사용하여 국가 간 공동연구 빈도 행렬을 생성하였다. 논문 A에서 총 저자가 다섯 명이고 각 저자 소속 기관의 국가가 {China, China, China, Korea, Japan}인 경우 이로부터 유일한 국가를 추출하면 {China, Korea, Japan}이 된다. 이를 바탕으로 가능한 협업 조합인 {China, Japan}, {China, Korea}, {Japan, Korea} 세 가지 경우에 대해 협업 횟수를 1씩 증가시키는 방식을 사용하였다. 본 연구에서 수집된 총 6,634개의 연구논문에서 두 개 이상의 국가가 협업하여 출판한 연구논문은 총 943편으로써 전체의 14.2%를 차지하며 하나 이상의 국가와 협업을 수행한 국가는 총 79개로 나타났다.

<Fig. 3.4>의 네트워크에서 노드 크기는 연결 중심성에 비례하며 호(edge)의 굵기는 국가별 협업 연구논문의 수에 비례한다. 네트워크 시각화 측면에서 2회 이상 협업한 경우만 호가 생성되도록 설정하였다. 절대적인 연구논문 출판 수가 많은 China, United States, United Kingdom, Italy, Spain과 같은 국가들은 연결 중심성이 높은 국가들임을 알 수 있다. 국가 간 협업이 가장 활발한 경우는 China와 United States이며 지역 및 역사적 특성으로 인해 China와 Hong Kong의 협업 빈도 또한 상당히 높게 나타나는 것을 알 수 있다. 또한 정치적인 환경으로 인해 Taiwan은



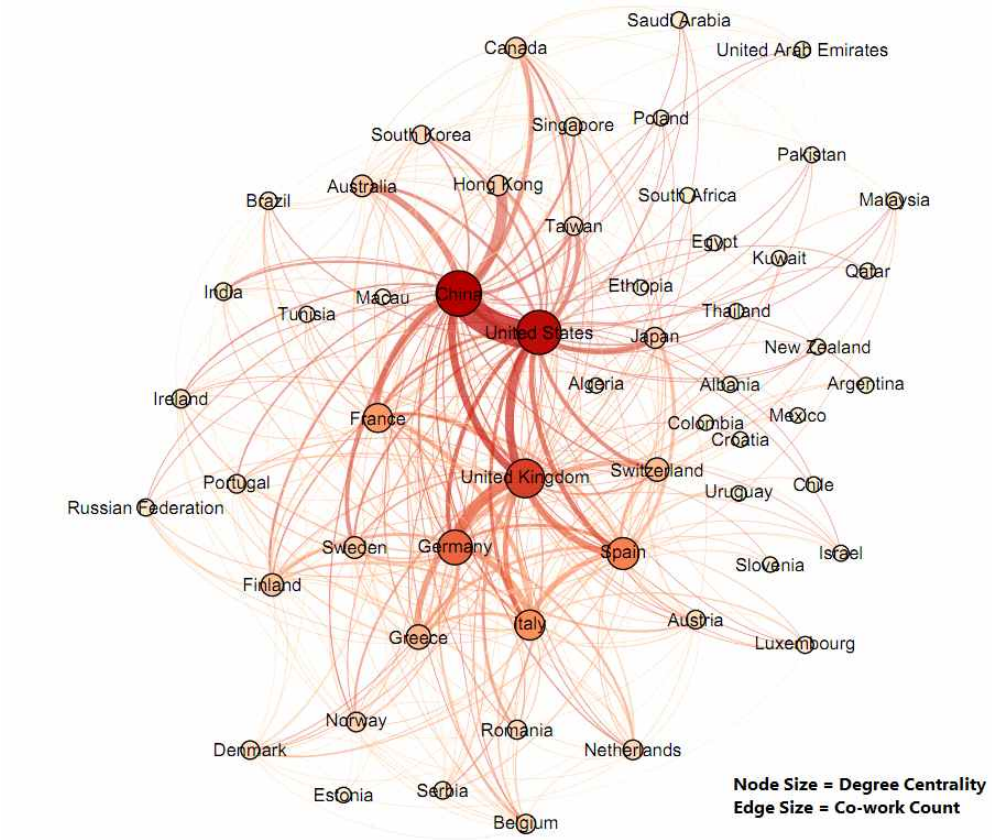
	China	USA	Russia	Korea	Japan
China	30	0	4	10	3
USA	0	10	6	2	6
Russia	4	6	3	7	8
Korea	10	2	7	4	6
Japan	3	6	8	6	6

<Fig. 3.3> 협업 네트워크를 위한 Symmetric matrix 생성 로직

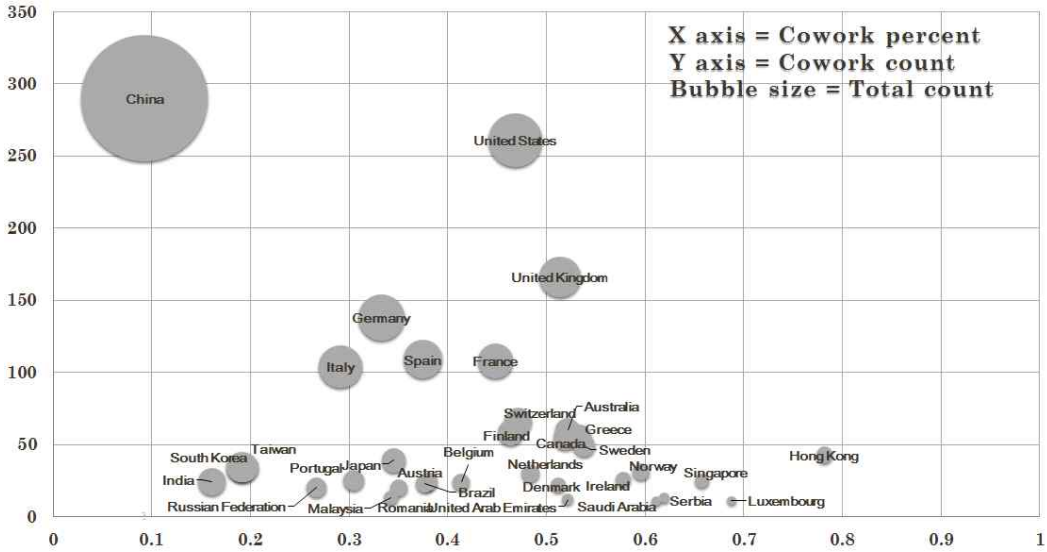
China보다 United States와 보다 많은 협업을 하는 것으로 나타났다. 지정학적인 위치 관계를 바탕으로 살펴보면 유럽의 경우에는 국가 간 협업이 상당히 활발한 것을 확인할 수 있다. United Kingdom, Germany, Spain, Italy, France를 중심으로 다른 유럽 국가들 사이의 연결 강도가 높은 경우가 빈번하게 나타난다.

이에 반하여 동북아 3국(Korea, China, Japan)의 경우 United State와는 협업의 빈도가 높게 나타나는 반면 세 국가 사이의 협업 빈도는 상대적으로 낮게 나타나는데, 이는 동북아 3국의 지정학적, 역사적 관계와도 무관하지 않은 것으로 판단된다.

<Fig. 3.5>는 최소 10편 이상을 출간한 국가들에 대해서 국가별 총 연구논문 수(버블의 크기), 협업 논문 수(y축), 및 국가 간 협업 비율(협업 논문 수/전체 논문 수, x축)을 나타낸 것이다. 앞에서 나타난 바와 같이 China는 총 출간 논문 수가 다른 국가에 비해 압도적으로 높은 반면 논문 수 대비 타 국가와의 협업 비율은 약 9.3%로 굉장히 낮은 수준인 것을 알 수 있다. 반면 연구논문 수 기준 2위를 차지하는 United States의 경우 협업 비율이 약 47%로서 China의 협업 비율 보다 다섯 배 이상 높게 나타나는 것을 확인할 수 있다. 또한 총 논문 수 대비 3위~7위를 차지하는 Germany, Italy, United Kingdom, Spain, France의 경우 약 30~50% 정도



<Fig. 3.4> Collaboration network among countries



<Fig. 3.5> The number of published papers (bubble size), country-level collaborated papers (y-axis), and collaboration ratio (x-axis)

의 협업 비율을 나타내고 있다. 한국의 경우 협업 비율이 20% 미만으로써 대상 국가 중 하위권에 속하는 것을 알 수 있는데 이는 사물인터넷 관련 연구가 국내 연구진들 위주로 진행되고 있음을 나타내는 것으로써 국가 간 협업에 보다 적극적으로 참여할 필요가 있음을 확인할 수 있다. 협업 네트워크 및 빈도 분석을 통해 알 수 있는 특이한 사항은 도시 국가인 Singapore와 Hong Kong의 경우 국가간 협업 비율이 매우 높게 나타난다는 것이다. Hong Kong의 경우 80%에 근접하는 협업 비율을 나타내고 있으며 Singapore의 경우도 65% 수준의 비율을 나타내는데, 이는 다양한 산업 분야에서 허브 역할을 하는 두 국가의 특성이 학술 연구분야에도 유사하게 나타나는 것이라고 볼 수 있다.나라별 협업 출간수의 순위는 [Table 3.2]에 나타나 있다.

[Table 3.2] 나라별 협업 출간수 기준 상위 20위까지의 표

순위	나라1	나라2	협업횟수	순위	나라1	나라2	협업횟수
[1]	USA	China	92회	[11]	Italy	Spain	19회
[2]	Germany	UK	39회	[12]	Switzerland	Germany	18회
[3]	Hong Kong	China	37회	[13]	Canada	China	17회
[4]	USA	UK	29회	[14]	Italy	UK	17회
[5]	UK	China	25회	[15]	Taiwan	USA	17회
[6]	Spain	UK	24회	[16]	USA	Spain	17회
[7]	Australia	China	23회	[17]	Japan	China	16회
[8]	Greece	UK	22회	[18]	France	Italy	16회
[9]	Germany	Italy	22회	[19]	China	Japan	16회
[10]	China	France	20회	[20]	Germany	USA	15회

3.3. 네트워크 중심성 분석

협업 네트워크상에서 중요한 역할을 하는 국가를 판별하기 위하여 본 연구에서는 연결 중심성과 매개 중심성 지표를 사용하였다. <Fig. 3.3>에서 호의 가중치(weight)는 두 국가 간의 협업 횟수로 설정이 되어 있으므로 이를 그대로 사용하여 연결 중심성을 산출할 수 있다. 그러나 두 노드 간의 최단 거리를 구하는 매개 중심성의 경우, 두 국가 간의 협업 횟수가 많을수록 두 국가의 거리가 가깝게 산출되어야 하므로 본 연구에서는 각 국가의 매개 중심성을 산출할 경우에는 호의 가중치를 <Eq. 3.1>과 같이 변환 하였고, 국가 간 협업 횟수가 존재하지 않을 경우에는 대수 1000을 가중치로 설정하여 사용하였는데, 이를 통해 경로가 절단되는 문제점을 보완하였다.

$$\begin{cases} C_{ij} = 0 \,,\, Edgeweight = 1000 \\ C_{ij} \neq 0 \,,\, Edgeweight = \dfrac{1}{C_{ij}} \end{cases}$$
$$C_{ij} = \text{국가} i \text{와 국가} j \text{사이의 협업 횟수}$$

<Eq. 3.1> Edge weight equation

[Table 3.3]과 [Table 3.4]는 국가 간 협업 네트워크상에서 연결 중심성 및 매개 중심성 상위 20개 국가들을 나타낸 표이다. 연결 중심성 측면에서는 상위 국가들은 협업논문 출판 수 순위와 거의 일치하나 논문 수의 차이에 비해 연결 중심성 차이는 상대적으로 적게 나타나는 것을 확인할 수 있다. 연결 중심성 순위 1위인 China의 경우 협업 논문 대비 연결 중심성의 수치가 1.21배로 나타나나, 2위인 United States는 1.28배, 3위와 4위인 United Kingdom과 Germany는 1.65배, 5위인 Spain은 1.79배로써 상대적으로 높게 나타나는 것을 확인할 수 있다. 매개 중심성 측면에서 보면 총 논문 및 협업 논문 수 기준 3위를 차지하는 United Kingdom이 가장 높은 매개 중심성을 나타내는 것으로 나타났다. 이는 United Kingdom은 절대적인 논문 수 기준으로는 1위와 2위인 China와 United States에 미치지 못하나 다양한 국가들과 폭넓은 협업 연구를 수행함으로써 연구의 중개자 역할을 하는 주요한 위치를 차지하는 국가라는 것을 확인시켜주는 결과라 할 수 있다. 또한, Belgium(18위 → 7위), Canada(14위 → 8위), Norway(17위 → 9위). Malaysia(32위 → 10위)등과 같은 국가들은 상대적으로 협업 횟수는 적으나 다양한 국가들과의 협업을 하고 있다는 것을 연결 중심성 대비 매개 중심성 순위 상승을 통해 확인할 수 있으며, 반대로 Greece(8위 → 13위), Switzerland(9위 → 20위), Sweden(11위 → 25위)과 같은 국가들은 국가 간 협업이 특정 국가에 편중되어 있어 연결 중심성 대비 매개 중심성 순위가 크게 하락하는 것을 볼 수 있다.

[Table 3.3] Top 20 countries for degree centrality

연결 중심성					
순위	국가	총 논문	협업 논문	협업 국가	연결 중심성
1	China	3131	290	37	351
2	United States	557	261	49	335
3	United Kingdom	323	166	42	274
4	Germany	415	138	34	228
5	Spain	291	109	39	195
6	Italy	357	104	31	174
7	France	241	108	37	160
8	Greece	105	56	23	109
9	Switzerland	140	66	26	98
10	Finland	125	58	23	86
11	Sweden	91	49	18	78
12	Australia	115	60	23	77
13	Japan	113	39	16	69
14	Canada	104	54	22	66
15	Hong Kong	55	43	15	63
16	Netherlands	62	30	21	56
17	Norway	52	31	19	53
18	Belgium	58	24	21	50
19	Romania	57	20	15	48
20	Austria	63	24	16	43

[Table 3.4] Top 20 countries for betweenness centrality

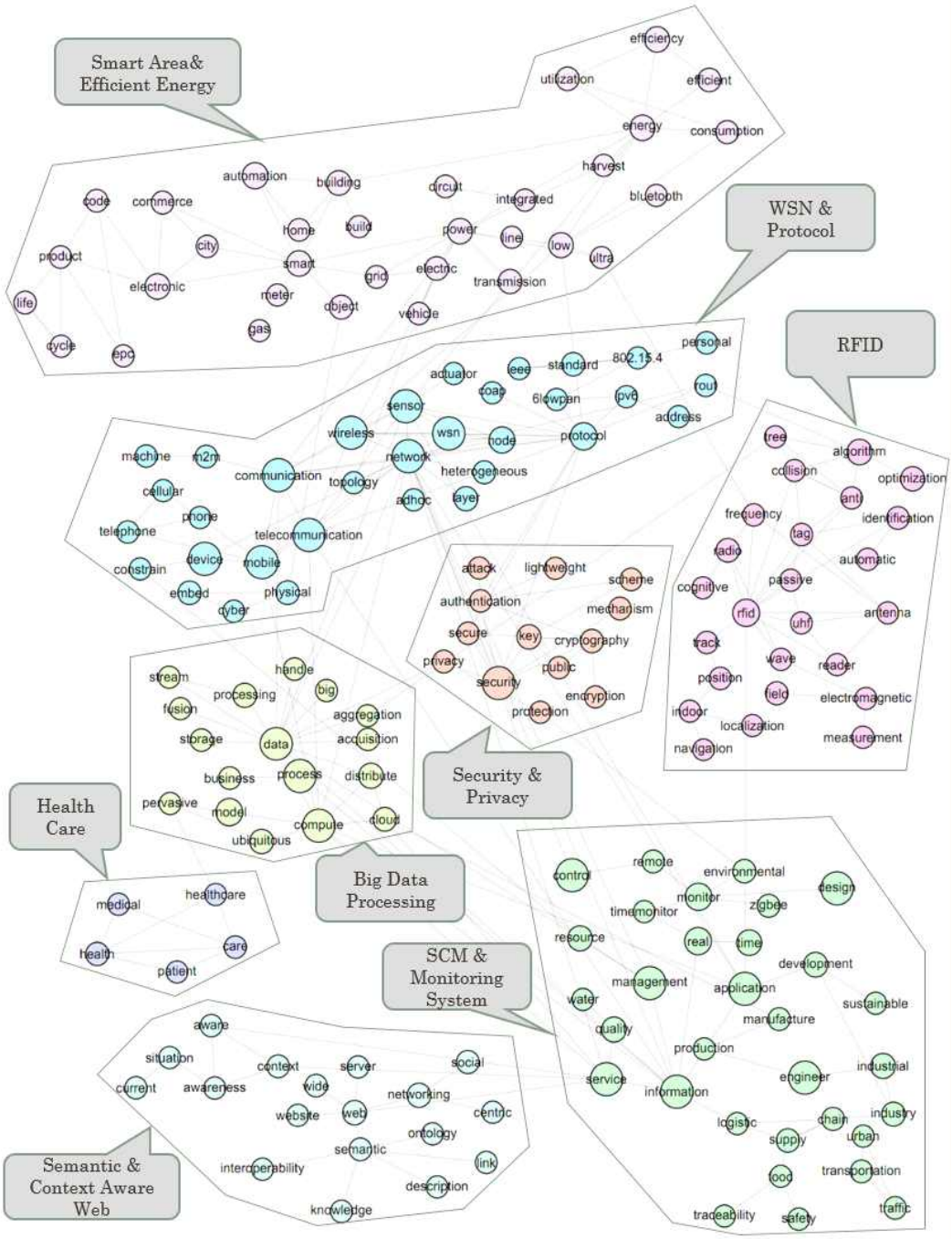
매개 중심성					
순위	국가	총 논문	협업 논문	협업 국가	매개 중심성
1	United Kingdom	323	166	42	0.4054
2	United States	557	261	49	0.3930
3	China	3131	290	37	0.3919
4	Germany	415	138	34	0.2043
5	Spain	291	109	39	0.1963
6	France	241	108	37	0.0943
7	Belgium	58	24	21	0.0509
8	Canada	104	54	22	0.0499
9	Norway	52	31	19	0.0496
10	Malaysia	38	13	14	0.0446
11	Italy	357	104	31	0.0356
12	Australia	115	60	23	0.0256
13	Greece	105	56	23	0.0256
14	Finland	125	58	23	0.0256
15	India	149	24	15	0.0256
16	South Korea	178	34	14	0.0256
17	Serbia	21	13	14	0.0256
18	Saudi Arabia	18	11	9	0.0256
19	Japan	113	39	16	0.0143
20	Switzerland	140	66	26	0.0073

3.4. 제목과 저자 키워드를 이용한 단어 네트워크 기반의 주요 연구분야 판별

논문의 주제를 판별하는 데 있어 가장 핵심적인 부분은 제목과 저자가 직접 자유로운 문자열로 제공하는 키워드 집합이다. 이에 더하여 1문단 내외로 작성되는 초록에 연구의 핵심 내용들이 서술되는 경우가 대부분이다. 따라서 본 연구에서는 2단계에 걸쳐 '사물 인터넷' 분야의 주요 연구분야 및 주제를 판별하고 이들의 관계를 분석하였다. 첫 번째 단계에서는 제목과 저자 제공 키워드를 이용하여 단어 유사도 네트워크를 구축하고 네트워크상에서의 군집을 판별하였다. 각 연구논문은 제목과 저자 제공 키워드로 구성된 벡터로 표현이 되며 전체 문서 집합은 TF-IDF로 표현되는 문서-단어 행렬로 변환하였다. 문서-단어 행렬에서 단어 사이의 코사인 유사도를 산출하고 이를 호의 가중치로 하여 <Fig. 3.6>과 같은 단어 유사도 네트워크를 구축한 결과 두 개의 기반 통신 기술 분야(WSN & Protocol, RFID), 세 개의 시스템 운영 분야(Semantic & Context Aware Web, Big Data Processing, SCM & Monitoring System), 세 개의 응용 산업 분야(Smart Area & Efficient Energy, Healthcare, Security & Privacy) 등 총 여덟 가지의 연구분야로 단어들이 군집화가 되는 것을 확인하였다.

기반 통신 기술 분야 중 WSN (Wireless Sensor Network) & Protocol과 관련해서는 Adhoc 네트워크 및 이를 위해 기기의 개별적 주소 할당을 가능케 하는 프로토콜인 IPv6 (Internet Protocol version 6), 이종 센서 네트워크와 IPv6 네트워크를 직접 연동하는 기술인 6LoWPAN (IPv6 Low-power Wireless Personal Area Networks), IEEE standard 802.15.4, CoAP (Constrained Application Protocol), 사물간 이동통신(M2M communication), 모바일 통신 관련 기기 등이 주요 단어로 판별되었다. 또 다른 통신 기술 분야인 RFID와 관련해서는 UHF (Ultra High Frequency) RFID Tag, Passive RFID, UHF대역 안에서 Antenna를 이용한 RFID, RFID 충돌방지 알고리즘(Anti Collision Algorithm), RFID를 이용한 위치추적(Track Position, Navigation, Indoor localization)등의 기술이 주로 연구됨을 확인할 수 있다.

마지막으로 응용 산업 분야와 관련해서는 최근 학계와 산업계에서 높은 관심을 갖고 있는 Healthcare 분야는 5개의 Medical, Patient, Care, Health, Healthcare 키워드가 Clique 구조를 지니고 있어 다른 분야들과는 달리 타 분야와의 연결성이 상당이 낮은 독립적인 연구분야라는 것을 알 수 있다. 반면 Security & Privacy 분야의 경우에는 기반 통신 기술 분야 및 시스템 운영 분야와 매우 밀접한 관계가 있음을 알 수 있으며 주요 연구 키워드로는 Lightweight authentication, Public-key Cryptography, Authentication scheme, Security mechanism등이 있는 것으로 나타났다. 또 다른 응용 산업 분야인 Smart Area & Efficient Energy에서는 Smart grid, Smart meter, Electronic

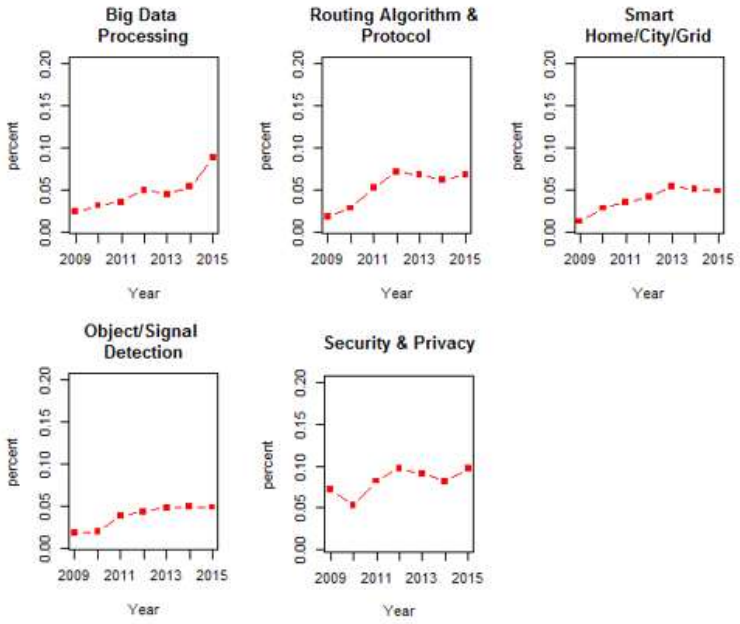


<Fig. 3.6> Keyword network based on "Title + Author provided keywords"

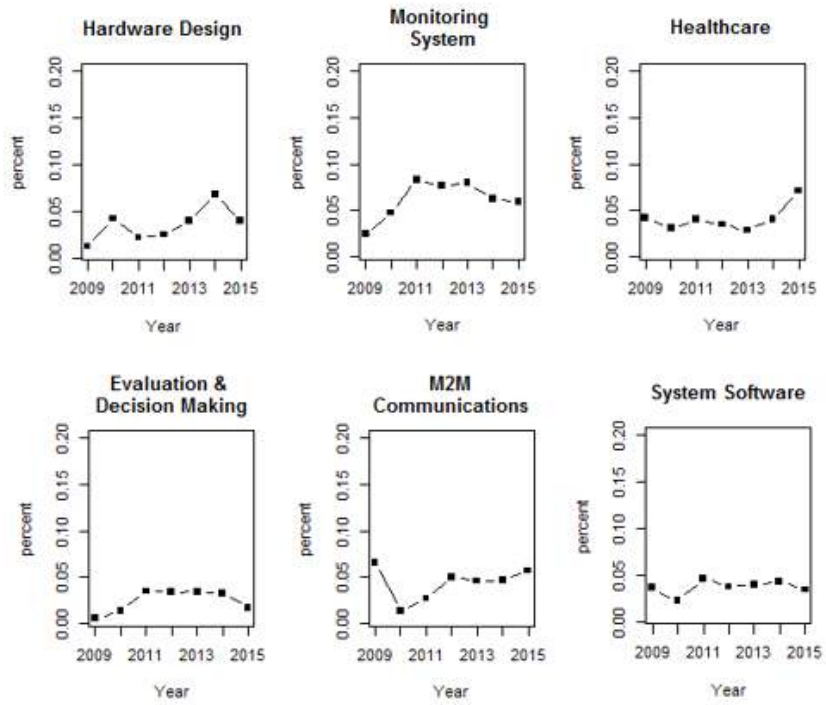
vehicle, Smart Commerce, Smart Home, Automation Building, Smart City 등 Smart Area 관련 키워드들과 Energy utilization, Energy consumption, Energy efficient, Energy harvest, Ultra low power integrated circuit, Low power transmission 등 에너지 효율성에 대한 연구 키워드가 공존하고 있는 것을 확인할 수 있다.

3.5. LDA를 이용한 토픽 추출 및 트렌드 분석

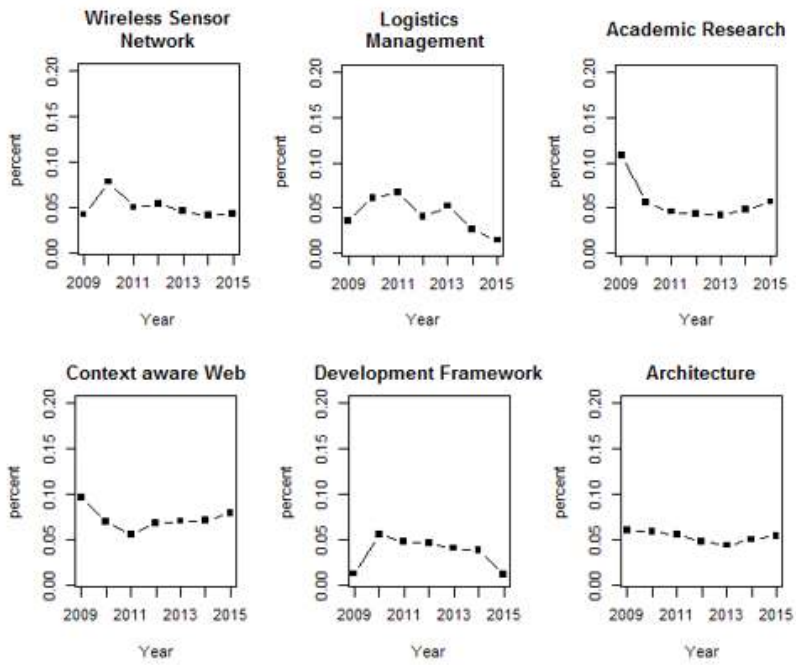
‘사물 인터넷’ 분야의 주요 연구분야 및 주제를 판별하기 위한 두 번째 단계로써 본 연구에서는 각 연구논문마다 제목, 초록 및 저자 제공 키워드를 결합하여 하나의 문서를 생성한 뒤 LDA를 이용하여 주제를 추출하였다. 수집된 논문 중 초록이 존재하지 않는 논문들을 제외하여 LDA 모델링에서는 총 6,595개의 연구논문을 분석 대상으로 사용하였다. LDA에서 토픽의 수는 20으로 설정하였으며 Collapsed Gibbs Sampling 기법을 통해 LDA 모델의 파라미터를 추정하였다. 또한 LDA에 의해 판별된 주제들의 시간에 따른 추세를 파악하기 위해 2009년 이전까지를 하나의 구간으로 설정하고, 2010년부터는 1년을 하나의 구간으로 설정하여 총 7개의 구간에서 각 주제가 전체 문서 집합에서 차지하는 비중의 변화를 통해 연구주제의 변화 패턴을 확인 하였다. 2009년 이전을 하나의 구간으로 설정한 이유는 <Fig. 3.1>에서 나타난 바와 같이 2003년부터 2009년까지의 연구논문 출간 수는 매우 적으며 2010년 이후로 그 수치가 급격히 증가하였기 때문이다. 개별 주제의 시간별 증가 혹은 감소 추세 패턴을 확인하기 위하여 본 연구에서는 각 주제가 전체 문서 집합에서 차지하는 비중이 종속변수이며 7개의 구간이 독립변수인 선형회귀분석 모형을 구축한 뒤, 회귀계수의 부호와 해당 회귀 계수의 유의확률에 따라 Hot/Warm/Cool/Cold 네 가지 유형으로 구분하였다[18]. LDA 모형에 의해 판별된 20개의 주제들의 시간에 따른 비중 변화는 <Fig. 3.7>에서 <Fig. 3.10>까지 나타나 있으며 Hot/Warm/Cool/Cold 유형으로 분류한 토픽의 결과는 [Table 3.5]와 같다.



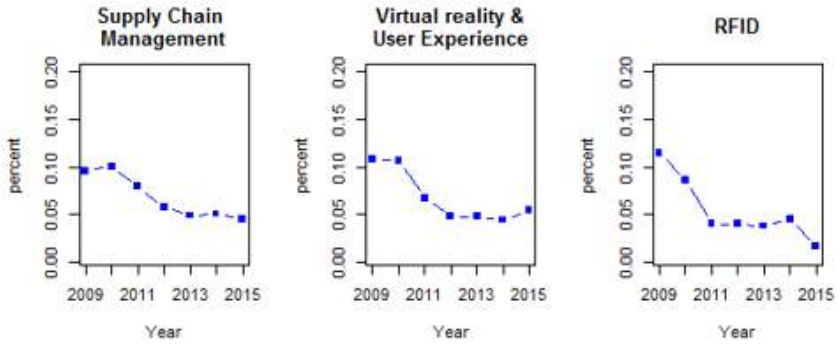
<Fig. 3.7> Hot Topics discovered by LDA and the changes of their proportions over time



<Fig. 3.8> Warm Topics discovered by LDA and the changes of their proportions over time



<Fig. 3.9> Cool Topics discovered by LDA and the changes of their proportions over time



<Fig. 3.10> Cold Topics discovered by LDA and the changes of their proportions over time

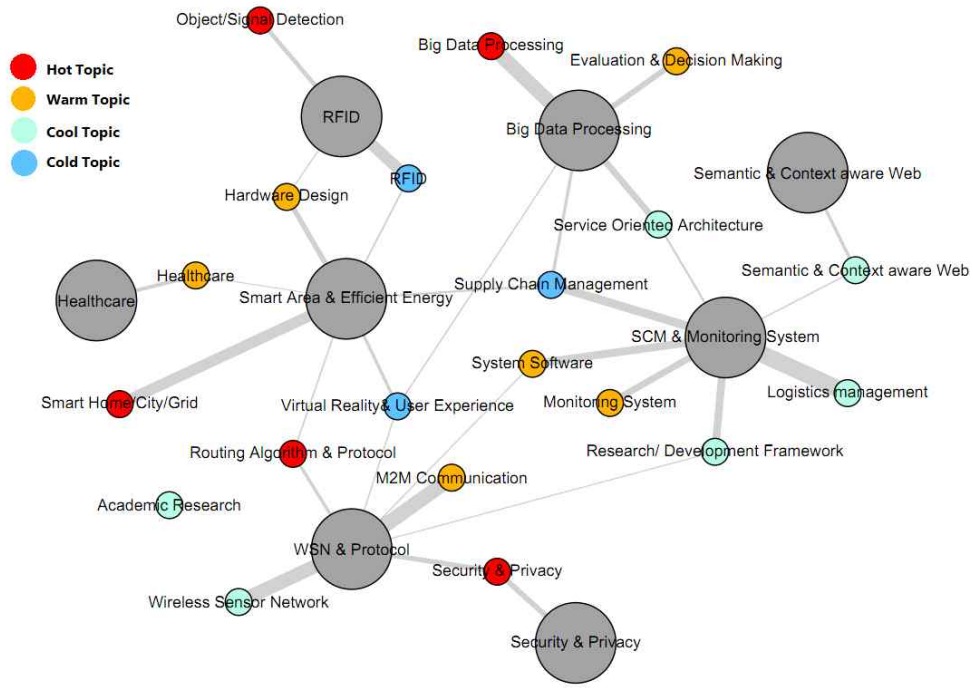
각 주제에 대해 발생 빈도가 높은 상위 30개의 단어는 부록에 제시되어 있다. LDA 모델링 대상 문서는 초록을 포함하였기 때문에 제목과 저자 키워드만으로 추출한 연구분야 <Fig. 3.6>과 비교할 경우 보다 구체적으로 나타나는 경우가 많은 것을 알 수 있다. 20개의 연구주제들 중에서 지속적으로 비중이 증가하고 있으며 그 증가분이 통계적으로 유의미한 Hot Topic 범주에는 총 다섯 개가 포함되었다. 최근 전세계적으로 각광을 받고 있는 Big Data Processing, Security & Privacy, Smart Home/City/Gird 관련 주제들이 사물인터넷 분야에도 핵심적인 연구분야로 자리 잡고 있음을 확인할 수 있으며, 기반 기술 중에는 Object/Signal Detection 및 Routing Algorithm & Protocol 이 꾸준하게 그 비중을 확장시켜 가고 있는 것으로 나타났다. 전체 대비 비중은 증가 추세에 있으나 그 증가분이 통계적으로 유의미하다고 판별되지는 못한 Warm Topic 범주에는 Healthcare, Monitoring System, M2M(Machine to Machine) Communication, System Software, Hardware design, Evaluation & Decision Making의 여섯 가지 주제가 포함되었다. 이 중 특이한 점은 Healthcare의 경우 Warm topic으로 분류되었으나 이는 2013년 이전 증가 폭이 크기 않아 나타난 결과이며 실제 2013년 이후로는 전체 20개의 주제 중 Big Data Processing과 더불어 그 비중이 가장 급격하게 증가하는 추세를 나타내는 주제라는 것이다. 전체 대비 비중이 감소 추세에 있으며 그 감소분이 통계적으로 유의미한 Cold Topic 범주에는 RFID, Virtual Reality & User Experience, 및 Supply Chain Management 세 가지 연구주제가 포함되었다. 세 가지 기술 모두 사물 인터넷 관련 기술의 핵심적인 요소라고 할 수 있으나 다른 연구주제들에 비해 상대적으로 연구가 먼저 시작된 주제들이기 때문에 전체 연구주제에서 차지하는 비중이 감소 추세에 있는 것으로 나타났다. 나머지 Semantic & Context aware Web, Logistics Management, Wireless Sensor Network, Research/Development Framework, Service Oriented Architecture, 및 Academic Research 주제는 전체 대비 비중이 감소 추세에 있으나 그 감소분이 통계적으로 유의미하지 않은 Cool

[Table 3.5] Identifying Hot/Warm/Cool/Cold Topics

Kind of topics	Topic description
Hot Topic (25%)	Security & Privacy Smart Home/City/Grid Big Data Processing Object/Signal Detection Routing Algorithm & Protocol
Warm Topic (30%)	Healthcare Monitoring System M2M Communications System Software Hardware design Evaluation & Decision Making
Cool Topic (30%)	Semantic & Context aware Web Logisitics Management Wireless Sensor Network Research/Development Framework Service Orientid Architecture Academic Research
Cold Topic (15%)	RFID Virtual reality & User Experience Supply Chain Management

Topic 범주로 분류되엇다. 이 중에서 Semantic & Context aware Web과 Service Oriented Architecture의 경우 회귀계수가 음수이기는 하나 전체 기간에 걸쳐 거의 일정한 수준의 연구 비중을 차지하는 주제라는 것을 그래프 상으로 확인할 수 있다. 한 가지 유의할 사항은 전술한 네 가지 범주는 상대적인 연구주제의 비중을 나타낸 것이기 때문에 Hot Topic에 속한 주제가 우수한 주제이고 Cold Topic에 속한 주제가 열등한 주제라는 결론을 내릴 수는 없다는 것이다. 그보다는 시간에 따른 연구주제의 트렌드 분석을 통해 과거 어떤 주제들이 상대적으로 빈번히 연구되었으며 최근에는 사물인터넷 관련 주요 연구분야의 비중이 어떻게 변화하였는지를 해석하는 데 적절히 사용되어야 할 것이다.

<Fig. 3.11>는 <Fig. 3.6>에서 단어들의 네트워크로 도출한 여덟 개의 사물인터넷 핵심 연구분야와 LDA를 통해 판별한 20개의 연구주제들 사이의 관계를 네트워크로 표현한 것이다. 회색 큰 원으로 표기된 마디는 제목과 저자 제공 키워드를 통해 도출한 주요 연구분야이며, 작은 원들은 LDA를 통해 판별한 연구주제들이다. 호의 굵기는 여덟 개의 핵심 연구분야에 포함된 단어들이 LDA를 통해 생성된 20개의 연구주제에서 차지하는 단어 확률값의 합과 비례한



<Fig. 3.11> Network graph constructed by combining the eight major research areas discovered by the keyword similarity network and 20 topics discovered by LDA.

다. <Fig. 3.11>를 통해 다음과 같은 현황을 파악할 수 있다. 첫째, 기반 기술과 관련된 WSN & Protocol과 시스템 운영과 관련된 SCM & Monitoring 및 Big Data Processing 분야는 LDA에서 판별된 20개의 주제 중 여러 개의 주제와 밀접하게 연관되어 있다는 것이다. 반면에 응용 분야에 속하는 Healthcare나 Security & Privacy의 경우 LDA에서 판별된 주제들 중 오직 하나의 주제와 연결되어 있는 독립적인 분야임을 알 수 있다. 다만 Smart Area & Efficient Energy의 경우 응용 분야 중에서도 그 적용 범위가 상당히 넓은 편이기 때문에 LDA에서 판별된 여러 개의 주제와 연관성이 높은 것으로 나타났다. 둘째, LDA에서 판별된 연구주제들의 비중 측면에서 볼 때, Cold Topic들은 대체적으로 여러 가지의 연구분야와 연관되어 있는 반면, Hot Topic들은 대부분이 하나의 연구분야에만 연결되어 있으며 Security & Privacy와 Routing Algorithm & Protocol만 각각 두 개의 연구분야와 연결이 되어 있는 것을 볼 수 있다. 진술한 바와 같이 Cold Topic에 속하는 주제들은 더 이상 중요한 연구주제가 아니라는 의미가 아니며 사물인터넷 관련 연구의 토대에 해당하는 주제들이라고 할 수 있다. 따라서 이러한 연구들은 상대적으로 초기에 활발하게 연구가 되었으며 그 연구 결과물들을 바탕으로 다양한 연구분야에서 새로운 연구가 수행되었음을 유추해볼 수 있다. 반면 Hot Topic과 관련된 주제들은 상대적으로 최근에 연구 비중이 높게 나타나는 주제들로써 아직은 각자 독자적인 분야로 연구가 되고 있는 것이라고 판단할 수 있다. Cold Topic에 해당되는 주제들과 마

찬가지로 시간이 지나게 되면 현재 Hot Topic으로 분류된 여섯 가지의 주제 역시 다양한 연구 결과물들을 바탕으로 여러 연구분야에 접목될 가능성이 높으며, 또 다른 새로운 연구주제(emerging topics)를 도출하는 시금석이 될 수 있을 것으로 기대한다.

Ⅳ. 결 론

본 연구에서는 2003년부터 2015년까지 출판된 국제 학술지 및 학술대회 논문을 바탕으로 사물인터넷 관련 연구의 동향을 분석하였다. 이를 위해

첫째, 연도-국가별 빈도 분석 및 국가간 협업 네트워크 분석을 통해 사물인터넷 분야에서 중요한 역할을 하는 국가들의 특징을 분석하였다.

둘째, 연구논문의 제목과 저자 제공 키워드를 바탕으로 단어간 유사성을 측정하여 총 여덟 가지의 주요 연구분야를 도출하고 각 분야에서 핵심적인 역할을 하는 단어들을 분석하였다.

셋째, 제목, 저자 제공 키워드 및 초록으로 구성된 문서들에 대해 LDA를 적용하여 20개의 주요 연구주제를 판별하고 해당 연구주제들의 핵심어 및 시간에 따른 연구 비중의 추이를 분석하였다.

마지막으로 단어간 유사성을 바탕으로 판별된 주요 연구분야와 LDA를 통해 도출된 주요 연구주제 사이의 관계를 네트워크로 표현하고 그 의미를 분석하였다.

본 연구의 결과물을 바탕으로 사물인터넷에 대하여 처음 접하는 사람이 동향과 핵심분야의 키워드를 쉽게 이해할 수 있도록 그래프와 분석결과를 산출 하였고, 더 나아가 사물인터넷 분야의 주요 연구주제 및 동향을 효율적으로 탐색할 수 있을 것이다, 본 연구는 사물인터넷의 전체적인 연구 트랜드를 분석하였다. 더 나아가서는 사물인터넷 특정분야의 대한 분석을 해당 방법론을 통해 연구 할 수 있을 것이며, 다른 분야에도 적용하여 유사한 결과물을 효과적으로 도출할 수 있을 것이다.

참고문헌

[1] Höller, J., Tsiatsis, V., Mulligan, C., Karnouskos, S., Avesand, S., and Boyle, D. (2014). **From Machine-to-Machine to the Internet of Things: Introduction to a New Age of Intelligence**, Elsevier.

[2] Manyika, J., Chuli, M., Bughin, J., Dobbs, R., Bisson, P., and Marrs, A. (20 13). Disruptive technologies: Advances that will transform life, business, an d the global economy, *McKinsey Global Institute*.

[3]Betsy, B. and Mike J. W. (2015). Hyper Cycle for Emerging Technologies, *Gartner*.

[4] Atzori, L., Iera, A., and Morabito, G. (2010). The internet of things: A sur vey, *Computer Networks*, **54**(15), 2787-2805.

[5] Xu, L., He, W., and Li, C. (2014). Internet of Things in Industries: A Surv ey, *IEEE Transactions on Industrial Informatics*, **10**(4), 2233-2243.

[6] Whitmore, A., Agarwal, A., and Xu, L. (2015). The Internet of Things - A survey of topics and trends, *Information Systems Frontiers*, **17**(2), 261-274.

[7] Otte, E. and Ronald, R. (2002). Social network analysis: A powerful strate gy, also for the information sciences, *Journal of Information Science*, **28**(6), 441-453.

[8] Newman, M. (2010). **Networks: An Introduction,** Oxford University Press.

[9] Freeman, L. C. (1978). Centrality in social networks conceptual clarificatio n, *Social Networks*, **1**(3), 215-239.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, **101**, 5228-5235.

[10] Park, H. W. and Leydesdorff, L. (2013). Decomposing social and semantic networks in emerging "big data" research, *Journal of Informetrics*, **7**(3), 756-765.

[11]Choi, J. H. and Hwang, Y. S. (2014). Patent keyword network analysis for improving technology development efficiency, *Technological Forecasting an d Social Change*, **83**, 170-182.

[12] Song, M., Go, E. H., and Lee, D. H. (2015). Identifying the landscape of Alzheimer's disease research with network and content analysis, *Scientom etrics*, **102**(1), 905-927.

[13] Blei, D. M. (2012). Probabilistic topic models, *Communications of the AC M*, **55**(4), 77-84.

[14] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocatio n, *Journal of Machine Learning Research* **3**, 993-1022.

[15] Song, M. and Kim, S. Y. (2013). Detecting the knowledge structure of bi oinformatics by mining full-text collections, *Scientometrics*, **96**(1), 183-201.

[16] Priva, U. C. and Asuterweil, J. L. (2015). Analyzing the history of cogniti on using topic models, *The Changing Face of Cognition*, **135**, 4-9.

[17] Kim, S. Y., Song, S. J., and Song, M. (2015). Investigation of topic trends in computer and information science by text mining techniques: From the perspe ctive of conferences in DBLP, *Korean Society for Information Management*, 3 2(1), 135-152.

[18] Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceed ings of the National Academy of Sciences*, **101**, 5228-5235.

Abstract

Research Trends Analysis for ‘Internet of Things’ based on Topic Modeling and Social Network Analysis

Kim, Jun Hong
(Supervisor Lee, Won Young)
Dept. of Data Science
The Graduate School of Seoul National University of Science and Technology

In this paper, we analyze the research trend related to “Internet of Things (IoT)” based on network analysis and topic modeling techniques. Based on a total of 6,634 articles that have been published in international journals and conferences, a publication year-country frequency analysis and a country-level collaboration analysis based on the degree and betweenness centralities are conducted. Then, eight major research areas are determined by the word similarity analysis based on the title and author provided keywords. In addition, we discover 20 research topics based on the title, author provided keywords, and abstract by employing the latent dirichlet allocation (LDA). Finally the relationship among the eight research areas and the 20 research topics are exploited. Sensing and communication network-related techniques are found to be fundamental basis, whereas big data processing, health care, and smart home/building/city are found to be emerging applications for IoT related research.

부록 1: Topic Modeling Result (Topic 1 ~ Topic 10)

Big Data Processing	Routing Algorithm & Protocol	Smart Home/ City/Grid	Object/Signal Detection	Hardware Design	Security & Privacy	Monitoring System	Healthcare	Evaluation & Decision Making	M2M Communications
data	network	smart	algorithm	power	security	monitor	health	model	device
process	node	energy	detection	low	protocol	system	system	dynamic	communication
storage	performance	home	location	antenna	privacy	environmental	life	analysis	mobile
query	simulation	building	position	design	key	equipment	people	evaluation	standard
big	rout	city	signal	signal	authentication	traffic	medical	decision	application
large	energy	grid	localization	circuit	scheme	vehicle	provide	performance	connect
processing	transmission	automation	optimization	transmission	secure	remote	technology	process	solution
stream	protocol	environment	technique	measurement	attack	realtime	care	factor	m2m
distribute	algorithm	infrastructure	improve	technique	trust	safety	live	behavior	machine
collect	time	power	recognition	material	access	condition	healthcare	characteristic	networking
amount	scheme	electronic	time	small	ipv6	control	patient	problem	telecommunication
database	reduce	appliance	accuracy	optical	mechanism	water	activity	risk	address
realtime	packet	consumption	image	electric	constrain	mine	increase	state	challenge
efficient	access	efficiency	experiment	harvest	6lowpan	environment	human	approach	access
search	consumption	provide	pattern	fiber	coap	maintenance	body	situation	mobility
time	channel	light	experimental	operation	cryptography	monitoring	personal	mechanism	capability
handle	mechanism	solution	error	line	identity	fault	develop	largescale	phone
analysis	delay	application	detect	receive	lightweight	temperature	individual	analyze	connectivity
mine	efficient	meter	feature	battery	requirement	parameter	daily	fuzzy	machinetomachine
generate	cost	build	indoor	electromagnetic	protection	operation	public	adopt	global
video	resource	give	target	synchronization	layer	emergency	electronic	effective	requirement
spatial	rate	experimental	estimation	chip	address	time	hospital	hierarchical	smartphone
acquisition	efficiency	facility	filter	performance	public	transmission	critical	rule	multimedia
cluster	schedule	efficient	measurement	integrate	encryption	warn	improve	assessment	type
aggregation	allocation	commerce	problem	energy	personal	disaster	include	strategy	bluetooth
engine	compare	block	distance	voltage	threat	coal	wearable	map	communicate
fusion	strategy	utilization	track	electronic	analysis	early	relate	give	wifi
store	quality	electric	accurate	field	computation	realize	disease	construct	connection
technique	optimization	apply	classification	sense	transfer	gas	group	prediction	require
scale	optimal	intelligence	environment	component	provide	real	assist	collaboration	provide

부록 2: Topic Modeling Result (Topic 11 ~ Topic 20)

System Software	Semantic & Context aware Web	Research/ Development Framework	Service Oriented Architecture	Wireless Sensor Network	Logistics management	Academic Research	Supply Chain Management	Virtual Reality & User Experience	RFID
system	web	technology	service	network	information	challenge	process	object	rfid
design	application	application	architecture	sensor	management	engineer	chain	user	identification
control	semantic	development	compute	wireless	system	potential	industry	social	tag
software	framework	layer	cloud	node	logistic	current	supply	physical	system
embed	approach	key	resource	wsn	technology	technological	product	compute	code
hardware	context	problem	platform	sense	integration	learn	manufacture	environment	reader
program	middleware	analyze	distribute	gateway	transportation	opportunity	business	ubiquitous	track
platform	heterogeneous	construction	provide	telecommunication	platform	international	industrial	interaction	electronic
module	interoperability	structure	infrastructure	communication	transport	survey	production	digital	epc
implement	domain	perception	discovery	zigbee	park	experience	enterprise	virtual	efficiency
interface	agent	field	composition	actuator	share	community	agricultural	human	technology
server	knowledge	publication	requirement	heterogeneous	operation	perspective	quality	space	time
operate	integration	rapid	quality	deployment	realize	innovation	agriculture	interface	identify
computer	provide	technical	provider	convergence	improve	learning	food	real	passive
requirement	ontology	basic	paradigm	integration	integrated	science	technology	online	application
remote	environment	trend	component	wsns	warehouse	recent	development	reality	read
component	integrate	apply	provision	testbed	integrate	ecosystem	material	provide	standard
zigbee	capability	core	virtualization	deploy	analysis	develop	market	pervasive	anticollision
controller	wide	relate	scalability	topology	distribution	explore	improve	networking	automatic
traditional	entity	principle	implement	wide	collection	approach	cost	memory	collision
prototype	describe	solve	solution	standard	publication	education	economic	interact	item
access	description	detail	functional	communicate	urban	major	traceability	vision	problem
terminal	reason	forward	scalable	configuration	railway	review	company	content	active
develop	contextaware	mode	integrate	local	realtime	practice	trace	form	reduce
realize	language	science	serviceoriented	802154	effective	aspect	customer	create	card
cyberphysical	link	point	offer	consist	analyze	evolution	engineer	computer	field
instrument	functionality	put	compose	multiple	exchange	advance	growth	paradigm	advantage
processor	representation	wide	share	lowpower	framework	offer	plan	prototype	slot
local	realworld	characteristic	element	include	realization	topic	modern	connect	frame
build	develop	campus	access	variety	modern	society	cycle	intelligence	wide